Data Collection and Preprocessing Phase

Date	12 June 2025
Team ID	SWTID1749627644
Project Title	Human Resource Management: Predicting Employee Promotions using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

Data Collection Plan & Raw Data Sources Identification Report outlines the approach for acquiring and identifying essential data needed for the project. It details the sources of raw data, the types of data collected, and ensures the data's suitability and quality to support accurate analysis and model building.

Data Collection Plan:

Section	Description
Project Overview	The machine learning project aims to predict employee promotions based on individual attributes. Using a dataset with features such as performance metrics, tenure, skills, and feedback, the objective is to build a model that accurately classifies promotion eligibility, supporting effective and data-driven workforce management decisions.
Data Collection Plan	● Search for datasets related to employee promotions, performance reviews, and HR records. ● Prioritize datasets with comprehensive information on tenure, skills, performance metrics, feedback, and demographic attributes.

Raw Data Sources Identified	The raw data for this project is sourced from Kaggle, a widely used platform for data science datasets and competitions. The dataset includes employee information such as department, region, education level, gender, recruitment channel, performance ratings, and training scores. This curated data enables the development of machine learning models to predict promotion eligibility within an organization

Raw Data Sources

Source Name	Description	Location/URL	Format	Size	Access Permissions
Skill Wallet Dataset	The dataset comprises employee details (gender, education, department), performance metrics (previous year rating, KPIs met, awards won), training information (number of trainings, training scores), and promotion outcomes.	https://drive.google.com/file/d/1I4qAYPpk3pctlYScWqw0Du2JEYF-rY80/view and https://www.kaggle.com/datasets/arashnic/hr-ana/data?select=train.csv	CSV	3.7 MB	Public